Research plan – Gabor T. Marth

1. Background

The era of genome sequencing has produced a complete reference sequence for many living organisms, most notably our own species. Because 999 out of 1,000 bases are identical when a pair of human chromosomes is compared, this reference captures most of what is shared within the DNA of all human kind. Sequence differences represent genetic variations – heritable markers that emerge from DNA mutations. The genomic patterns of variation structure are determined both by random processes such as genetic drift and mitotic recombination modulated by the effects of demographic history, and non-random, evolutionary processes such as the various forms of selection. Genetic variation is important because phenotypic variety arising from genetic causes is nearly always caused by allelic difference. Therefore, the study of genetic variants can lead to a better understanding of the connection between sequence and function, and explain medically important phenotypic traits including those involved in predisposition to human disease.

Fueled by this promise, and enabled by technological advances in sequencing and genotyping, public and private discovery projects have identified millions of polymorphic sites in the human genome. Initially, methods to interpret variations were unprepared for the rapid influx of data. Existing methods of population genetics lack the generality to simultaneously explain genome-scale variation data from qualitatively different observation processes. The neutral theory of sequence variations contributes relevant predictions but most of these were derived under very restrictive, simplifying conditions, due to the mathematical difficulties involved, and for lack of large-scale data sets in which theoretical results can be experimentally tested. Evolutionary theories of selection are valid but the bulk of genome DNA is not under selective constraints. Finally, the current practice of pharmaco-genetics is largely unaware of even the most basic patterns of human variation structure, and lacks the right tools to bear on connecting phenotype to genotype. There is clearly a need for new theory in population genomics that enables us to build models that explain genome-wide data sets of experimental data, captures salient features of the variation landscape, and answers practical questions relevant both to experimental design and to the interpretation of experimental results.

2. Contribution to population genomics

As a post-doc at Washington University my main focus was on genome sequencing informatics. After being involved in the technical details of sequencing software it was a logical transition to polymorphism mining in redundant sequence data. I lead the development of the polymorphism discovery method PolyBayes, a collection of algorithms that take advantage of the genome reference sequence as a substrate to organize and multiply align redundant, fragmentary sequence data, employ a paralog discrimination procedure to screen out sequence duplicates, and a mathematically rigorous, Bayesian detection algorithm to identify polymorphic sites. This tool was used to mine single-nucleotide polymorphisms (SNPs) from ESTs aligned to the reference genome, and from whole-genome shotgun reads produced by The SNP Consortium. Together with my collaborators at the NCBI, I have applied this analytical tool to the analysis of the overlapping regions of large-insert (BAC) clones that were sequenced for the public human reference genome, and produced over 500,000 high quality candidate SNPs, and over 100,000 deletion-insertion type polymorphisms (DIPs). To complement these data, collaborators at Washington University and I have also produced sample-based allele frequency estimates in multiple ethnic populations for a large subset of these variations. Such a large data set has made it possible, for the first time, to explore the genomic distribution of human variability on the scale of the entire genome.

To interpret these observations, I have extended prior work in coalescent theory and derived new theoretical models of SNP marker density and allele frequency spectrum that take into account genetic drift, realistic values of recombination, and dynamic scenarios of population history. Using the BAC overlap SNP data to parameterize the models I predicted a significantly higher level of linkage disequilibrium, and a significantly reduced level of haplotype diversity within human populations than previously anticipated – predictions that have now been confirmed by experimental data from multiple laboratories. Additionally, I was able to show that the orthogonal data sets of marker density and allele frequency spectrum can be explained by the same population history, a finding that increases the confidence in these results. I am currently focusing on comparing the variation structure and the underlying molecular and demographic processes between geographically different world populations. Ascertainment of the commonality and difference between the haplotype structures of these populations is vitally important for the proposed comprehensive human haplotype map project – to ensure the generality of this expensive resource. Initial results based on models of population subdivision show that there are significant differences that need to be taken into account in the experimental design of both marker discovery and genotyping.

3. Proposed research program

My ultimate goal is to understand the fundamental forces that shape genetic variability. Random genetic drift, demographic history, the molecular mutation process, recombination, and the various forms of selection all act together in imprinting observable distributions of polymorphisms such as marker density, inter-marker spacing, allele frequency spectrum, and in determining human haplotype structure. The ability to quantify the effects of these forces will enable better detection of signals of direct medical importance painted upon a background of random genetic variation. The method of choice towards this goal is to develop dynamic models of these forces and to evaluate and parameterize each model by comparing its predictions to distributions observed in experimental data. Based upon these comparisons, models are then refined and re-parameterized. The resulting models can then be used to predict important characteristics of genome variation structure that are impossible or impractical to measure directly. These predictions not only help in the interpretation of the public variation resource, they also aid in the presentation of variation data to the public in a thoughtful, useful fashion that is of benefit to the research and the medical communities. Novel in this proposal is the integration of population genetics theory, the corresponding algorithmic design, as well as software implementation with the use of multi-locus, genome-wide data sets for hypothesis testing. The principal components of this project are described below.

Theory development: Building on existing results of population genetics, I intend to continue building quantitative models that describe the effects of the molecular and demographic forces that shape the observable distributions of genetic variations. The aim is to derive closed mathematical formulae whenever possible because these are, in general, easier to study, and faster to compute. Often, however, the processes can only be cast in terms of simulation procedures. In these cases, relevant distributions and summary quantities can still be obtained by tabulating large numbers of simulation replicates. In particular, I plan to focus on modeling the effects of non-uniform mutation rates, recombination hotspots and the process of gene conversion, as well as the effects of differential demographic history on the haplotype structure of human populations.

Unification of data from different mutation systems: The four main molecular systems currently used in variation studies are: micro satellite or simple tandem repeat polymorphisms (STRPs), mitochondrial variations, substitution type single-nucleotide polymorphisms (SNPs) and deletion/insertion polymorphisms (DIPs). While the mutation processes that produce these classes of variations differ in essential ways, the underlying demography is, by definition, the same. If one can demonstrate that the demographic conclusions based on these systems are fundamentally congruent, then these different variation types can be leveraged to bring different epochs of human demographic past into better focus.

Study of human haplotype structure to evaluate resources for medical use: A haplotype is an ordered set of allelic states, measured in a single strand of DNA, that are close enough in physical placement to be co-inherited as a single unit from generation to generation. Haplotypes have significantly higher information content as compared to the component genotypes, therefore carrying more resolving power for disease association studies. A recently proposed public "haplotype map" project is motivated by reports of experimental results showing a block structure consisting of local regions where haplotype diversity is greatly reduced. This means that, within such a region, the haplotype diversity of an entire population can be described by a small, minimal set of markers (enough to resolve the few different haplotypes observed), also reducing the total number of markers needed for a whole-genome association study. The true benefits of such a project, however, depend on many additional questions such as the selection of markers that define the haplotypes, the constancy of the blocks across parallel, independent sets of samples drawn from the same population, and across alternative definitions of block boundaries. One must also investigate whether current models of population history can account for the observed blocks or the theory needs to be revised, perhaps including in the models recombination hotspots and the effects of gene conversion to account for the breaks between blocks. Another important question is whether it is necessary to invoke selection, hitchhiking or other function-related phenomena to explain genome-wide haplotype structure even though most of the genome is unlikely to obey strong selective constraints. Once the block structure of a given population is successfully characterized, the question remains whether this structure is generalizable to all world populations, and whether one set of markers is sufficient to describe haplotype in all populations, or one must also include polymorphisms that are private to each population. Haplotypes are emerging as an important component of genome annotation, and these studies will also inform the genome annotation process with the latest results on the structure and application of haplotype data.

Study of allelic association with human disease: A successful haplotype project is expected to provide an effective tool for finding common alleles associated with complex diseases, as long as these alleles appear on a common haplotype background. In my opinion, it is unlikely that the majority of disease-causing alleles will fall into this category. Nevertheless, a thorough understanding of the shape of human variation landscape will help us to interpret local variation patterns, and inform us of the evolutionary origin of known polymorphisms that are the cause of functional allelic differences. When searching for unknown disease-causing alleles, whole-genome association scans using hundreds of thousands of markers simultaneously are likely to face serious theoretical-statistical challenges, as well as practical difficulties associated with the management of data sets of enormous size and complexity. Incorporation of additional facets of information should help overcome some of these problems. Lists of candidate genes thought to be involved in a given disease exist. These lists can possibly be extended by mining the disease literature with computational means, and by uncovering novel ontological relationships between genes on the basis of related phenotypes within model organisms. Focusing on the genomic regions that contain all the genes present even in a very inclusive candidate list is still likely to reduce the number of relevant markers to a number that is tractable for the detection of meaningful associations. Such practical considerations, when combined with haplotype data deployed in a way that utilizes our knowledge of genome variation structure should lead to novel methods in the experimental design for disease association studies.

Technology development: Prerequisite to the research described above is the development of efficient computational algorithms that allow one to study human haplotypes in a large number of samples, over long stretches of DNA sequence, under models of realistic complexity. This will require bringing the talents of both mathematicians and sophisticated programmers to bear on problems of computational complexity, speed, and numeric stability. The expected end-result is a suite of highly efficient computational and visualization tools, available in the public domain, to enable us, as well as other research groups to interpret variation data *vis-à-vis* phenotype either in local regions, or on the genome scale.

4. Teaching

Although my current appointment at the NCBI does not have a regular teaching component, I have been continuously involved in teaching since my graduate student years. At the School of Engineering at Washington University in St. Louis, I was responsible for the curriculum development and teaching of two upper undergraduate level courses in the Department of Systems Science and Mathematics: Process Control Laboratory, and Numerical Methods (Numerical Analysis). Additionally, I served as course consultant for a variety of other courses offered by other engineering departments. Since October 1999, I have been on the faculty of the Bioinformatics course at the Cold Spring Harbor Laboratory, and on the faculty of the Genomics course offered by the Canadian Bioinformatics Network. I have also been invited as guest lecturer at the Bioinformatics course offered by the National Institutes of Health. I have always enjoyed teaching, and found the interaction with students invigorating and intellectually refreshing. At the Genome Sequencing Center I had two fresh university graduates on my team helping me with various informatics projects. Although in a less formal setting, this was also an opportunity to educate (and be educated by) young people with razor-sharp minds. The opportunity of closer interaction with graduate and undergraduate students, both in the classroom, and as a graduate advisor is an important reason why I keenly look forward to a university position.